Module 10

Principal Component Analysis / Shrinkage

Kentaro Nakamura

GOV 2003

April 17th, 2026

Agenda

- Review of Midterm
- Principal Component Analysis (PCA)
- Shrinkage method
  - Ridge regression
  - Lasso regression

Midterm: Sensitivity Analysis (1)

- Consider the following data generating process:

$$\boldsymbol{Y} = \beta \boldsymbol{T} + \boldsymbol{X}\boldsymbol{\gamma} + \delta \boldsymbol{U} + \boldsymbol{\epsilon} \quad \mathbb{E}[\epsilon_i \mid \boldsymbol{T}, \boldsymbol{X}, \boldsymbol{U}] = 0$$

  - $\boldsymbol{U}$ is unobserved

- Now, suppose that you fit the regression above without $\boldsymbol{U}$. How does the bias look like?

- Now, using FWL theorem,

$$\hat{\beta} = (\boldsymbol{T}^\top (\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}}) \boldsymbol{T})^{-1} \boldsymbol{T}^\top (\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}}) \boldsymbol{Y} = \frac{\boldsymbol{T}^\top (\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}}) \boldsymbol{Y}}{\boldsymbol{T}^\top (\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}}) \boldsymbol{T}}$$

Midterm: Sensitivity Analysis (2)

- As we have $\boldsymbol{Y} = \beta \boldsymbol{T} + \boldsymbol{X}\gamma + \delta \boldsymbol{U} + \boldsymbol{\epsilon}$,

$$\begin{aligned}\hat{\beta} &= \frac{\boldsymbol{T}^\top (\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}}{\boldsymbol{T}^\top (\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{T}} \\ &= \beta + \frac{\boldsymbol{T}^\top (\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{X}}{\boldsymbol{T}^\top (\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{T}}\gamma + \delta \frac{\boldsymbol{T}^\top (\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{U}}{\boldsymbol{T}^\top (\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{T}} + \frac{\boldsymbol{T}^\top (\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{\epsilon}}{\boldsymbol{T}^\top (\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{T}}.\end{aligned}$$

- By the weak law of large numbers,

$$\begin{aligned}\frac{1}{n}\boldsymbol{T}^\top (\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{\epsilon} &\xrightarrow{p} \mathbb{E}[T_1 (I - \boldsymbol{P}_{\boldsymbol{X}_1})\epsilon_1] \\ &= \mathbb{E}[T_1 (I - \boldsymbol{P}_{\boldsymbol{X}_1})\mathbb{E}[\epsilon_i \mid \boldsymbol{T}, \boldsymbol{X}, \boldsymbol{U}]] = 0,\end{aligned}$$

- Similarly, the second term is

$$\frac{1}{n}\boldsymbol{T}^\top (\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{X} = 0.$$

and thus it is also negligible.

Midterm: Sensitivity Analysis (3)

- Finally, the middle term is written as

$$\delta \frac{\boldsymbol{T}^\top (\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}}) \boldsymbol{U}}{\boldsymbol{T}^\top (\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}}) \boldsymbol{T}} = \delta \times \frac{\frac{1}{n} \{(\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}}) \boldsymbol{T}\}^\top (\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}}) \boldsymbol{U}}{\frac{1}{n} \{(\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}}) \boldsymbol{T}\}^\top (\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}}) \boldsymbol{T}}.$$

- The numerator is

$$\frac{1}{n} \{(\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}}) \boldsymbol{T}\}^\top (\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}}) \boldsymbol{U} \xrightarrow{p} \mathbb{E}[T_i^{\perp \boldsymbol{X}} U_i^{\perp \boldsymbol{X}}] = \mathrm{Cov}(T_i^{\perp \boldsymbol{X}}, U_i^{\perp \boldsymbol{X}})$$

- The denominator is

$$\frac{1}{n} \{(\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}}) \boldsymbol{T}\}^\top (\mathbf{I}_n - \boldsymbol{P}_{\boldsymbol{X}}) \boldsymbol{T} \xrightarrow{p} \mathbb{E}[(T_i^{\perp \boldsymbol{X}})^2] = \mathbb{V}[T_i^{\perp \boldsymbol{X}}]$$

- Therefore, we get

$$\hat{\beta} \xrightarrow{p} \beta + \delta \times \frac{\mathrm{Cov}(T_i^{\perp \boldsymbol{X}}, U_i^{\perp \boldsymbol{X}})}{\mathbb{V}[T_i^{\perp \boldsymbol{X}}]}$$

Midterm: Sensitivity Analysis (4)

- Now, δ is the coefficient of unobserved confounder. So, by FWL,

$$\delta = \frac{\mathrm{Cov}(Y_i^{\perp \boldsymbol{X},T}, U_i^{\perp \boldsymbol{X},T})}{\mathbb{V}[U_i^{\perp \boldsymbol{X},T}]}.$$

- Also, recall that

$$\mathbb{V}(Y_i - \widehat{Y}_i) = (1 - R^2)\mathbb{V}(Y_i)$$

  - This is because $\mathbb{V}(Y_i) = \mathbb{V}(\widehat{Y}_i) + \mathbb{V}(Y_i - \widehat{Y}_i)$ and $R^2 = \frac{\mathbb{V}(\widehat{Y}_i)}{\mathbb{V}(Y_i)}$

- Thus,

$$\mathbb{V}[U_i^{\perp \boldsymbol{X},T}] = \mathbb{V}[U_i^{\perp \boldsymbol{X}}] \times (1 - R^2_{U \sim T|\boldsymbol{X}}) = \mathbb{V}[U_i^{\perp \boldsymbol{X}}] \times (1 - R^2_{T \sim U|\boldsymbol{X}})$$

  and we get

$$|\hat{\delta}| = \sqrt{R^2_{Y \sim U|T,\boldsymbol{X}}} \sqrt{\frac{\widehat{\mathbb{V}}(Y^{\perp \boldsymbol{X},T})}{\widehat{\mathbb{V}}(U^{\perp \boldsymbol{X}})(1 - R^2_{T \sim U|\boldsymbol{X}})}}.$$

Midterm: Sensitivity Analysis (5)

- Finally, notice that

$$R^2_{T\sim U|\boldsymbol{X}} = \mathrm{Corr}(T_i^{\perp\boldsymbol{X}}, U_i^{\perp\boldsymbol{X}})^2 = \frac{\mathrm{Cov}(T_i^{\perp\boldsymbol{X}}, U_i^{\perp\boldsymbol{X}})^2}{\mathbb{V}[T_i^{\perp\boldsymbol{X}}]\mathbb{V}[U_i^{\perp\boldsymbol{X}}]}$$

  and thus

$$\left|\frac{\mathrm{Cov}(T_i^{\perp\boldsymbol{X}}, U_i^{\perp\boldsymbol{X}})}{\mathbb{V}[T_i^{\perp\boldsymbol{X}}]}\right| = \sqrt{R^2_{T\sim U|\boldsymbol{X}}} \times \sqrt{\frac{\mathbb{V}[U_i^{\perp\boldsymbol{X}}]}{\mathbb{V}[T_i^{\perp\boldsymbol{X}}]}}.$$

- Therefore,

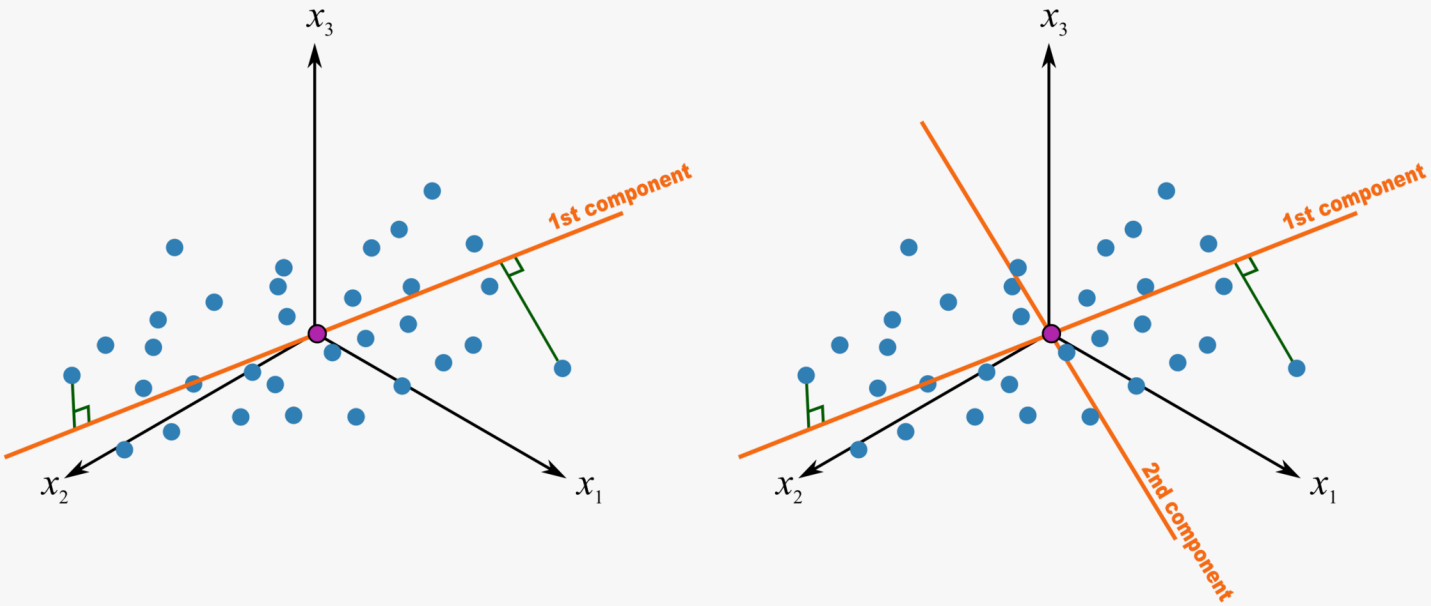
$$|\mathrm{bias}| = |\hat{\delta}| \times \left|\frac{\mathrm{Cov}(T_i^{\perp\boldsymbol{X}}, U_i^{\perp\boldsymbol{X}})}{\mathbb{V}[T_i^{\perp\boldsymbol{X}}]}\right|$$

$$= \sqrt{R^2_{Y\sim U|T,\boldsymbol{X}}}\sqrt{\frac{\mathbb{V}[Y_i^{\perp\boldsymbol{X},T}]}{\mathbb{V}[U_i^{\perp\boldsymbol{X}}] \times (1 - R^2_{T\sim U|\boldsymbol{X}})}}\sqrt{R^2_{T\sim U|\boldsymbol{X}}} \times \sqrt{\frac{\mathbb{V}[U_i^{\perp\boldsymbol{X}}]}{\mathbb{V}[T_i^{\perp\boldsymbol{X}}]}}$$

$$= \sqrt{R^2_{Y\sim U|T,\boldsymbol{X}}} \times \sqrt{\frac{R^2_{T\sim U|\boldsymbol{X}}}{1 - R^2_{T\sim U|\boldsymbol{X}}}} \times \sqrt{\frac{\mathbb{V}[Y_i^{\perp\boldsymbol{X},T}]}{\mathbb{V}[T_i^{\perp\boldsymbol{X}}]}}.$$

Principal Component Analysis (PCA)

- **Goal**: Dimension reduction
  - We often have so many variables, and we want to reduce the dimensionality and discover which dimension matters the most
  - We want to transform high-dimensional, correlated data, into a smaller set of uncorrelated variable (known as **principal component**)



- Notice that the first principal component is the direction where the variance is maximized.

Principal Component Analysis (PCA) (1)

- Let $x_1, \cdots, x_N \in \mathbb{R}^p$ be the observations, and let $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
  - We can then define the centered data matrix

$$\bar{X} = \begin{bmatrix} (x_1 - \bar{x})^\top \\ \vdots \\ (x_N - \bar{x})^\top \end{bmatrix} \in \mathbb{R}^{N \times p}$$

- The goal is that to find j-th principal component ($j \in \{1, \cdots, J\}$) that is the linear transformation of the data

$$s_j = \bar{X} a_j \in \mathbb{R}^N$$

where a_j is the coefficient

  - We want this transformation to satisfy (i) length is 1 and (ii) variance is maximized

Principal Component Analysis (PCA) (2)

- Let's calculate the variance of s_j. Now, as s_j's average is 0,

$$\mathrm{Var}[s_j] = \frac{1}{N} s_j^\top s_j = \frac{1}{N} a_j^\top \bar{X}^\top \bar{X} a_j = a_j^\top \mathrm{Var}[\bar{X}] a_j$$

- We want to find the projection vector a_j such that (i) length is 1 and (ii) variance is maximized. That optimization is written as

$$\max \quad a_j^\top \widehat{\mathrm{Var}[\bar{X}]} a_j \quad \text{such that} \quad a_j^\top a_j = 1$$

Principal Component Analysis (PCA) (3)

- By using Lagrangian multiplier, we have the Lagrangian

$$\mathcal{L}(a_j, \lambda) = a_j^\top \widehat{\mathrm{Var}[\bar{X}]} a_j - \lambda_j (a_j^\top a_j - 1)$$

and the derivative is

$$\partial_{a_j} \mathcal{L}(a_j, \lambda_j) = 2\widehat{\mathrm{Var}[\bar{X}]} a_j - 2\lambda_j a_j = 0$$
$$\partial_{\lambda} \mathcal{L}(a_j, \lambda_j) = a_j^\top a_j - 1 = 0$$

- Now notice that the first constraint is

$$\widehat{\mathrm{Var}[\bar{X}]} a_j = \lambda_j a_j$$

  which is eigenvalue problem!
  - Lagrangian multiplier λ_j becomes eignevalue
  - Note that we need to scale the eigenvector so that its length becomes 1.

Principal Component Analysis (PCA) (4)

- As a result, PCA corresponds to solving the eigenvalue decomposition of variance matrix $\widehat{\mathrm{Var}[\bar{X}]}$;
  - $\widehat{\mathrm{Var}[\bar{X}]} = \frac{1}{N}\bar{X}^\top \bar{X}$ is symmetric variance, so spectoral theorem guarantees the existence of EVD

- Moreover, the variance of j-th principal component s_j is

$$\widehat{\mathrm{Var}[s_j]} = \frac{1}{N}(\bar{X}a_j)^\top(\bar{X}a_j) = a_j^\top \widehat{\mathrm{Var}[\bar{X}]}a_j$$

and thus by eigenvalue decomposition $\widehat{\mathrm{Var}[\bar{X}]} = A\Sigma A^\top$ with $A = [a_1, \cdots, a_p]$ and $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_p)$, so

$$\widehat{\mathrm{Var}[s_j]} = a_j^\top A\Lambda A^\top a_j = e_j^\top \Lambda e_j = \lambda_j$$

  - That is, the eigenvalue corresponds to the variance of principal component

Principal Component Analysis (PCA) (5)

- Define the total variance of cenetred regressor as

$$\text{Total variance} := \sum_{j=1}^{p} \mathrm{Var}(X_j)$$

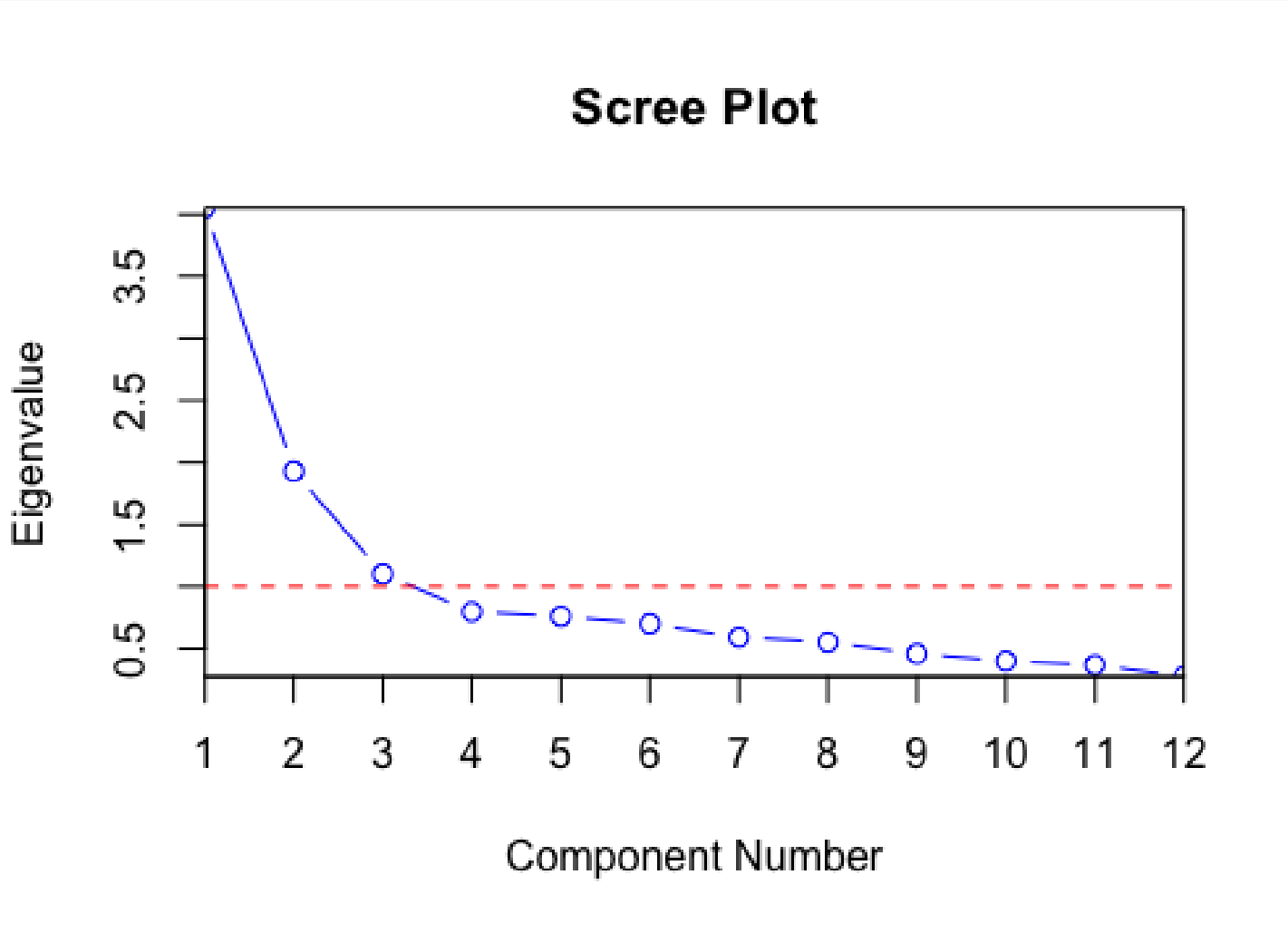
  - Now, notice that

$$\text{Total variance} = \mathrm{tr}(\widehat{\mathrm{Var}[\bar{X}]}) = \mathrm{tr}(A\Lambda A^\top) = \mathrm{tr}(\Lambda) = \sum_{j=1}^{p} \lambda_j$$

- How many principal components should we use?
  - If we use all p components (same number as the number of variables), we do not reduce the dimensions, and we can explain all the variations in the data
  - However, if we only use the very small number of principal components, we might not explain the variation in the data well.

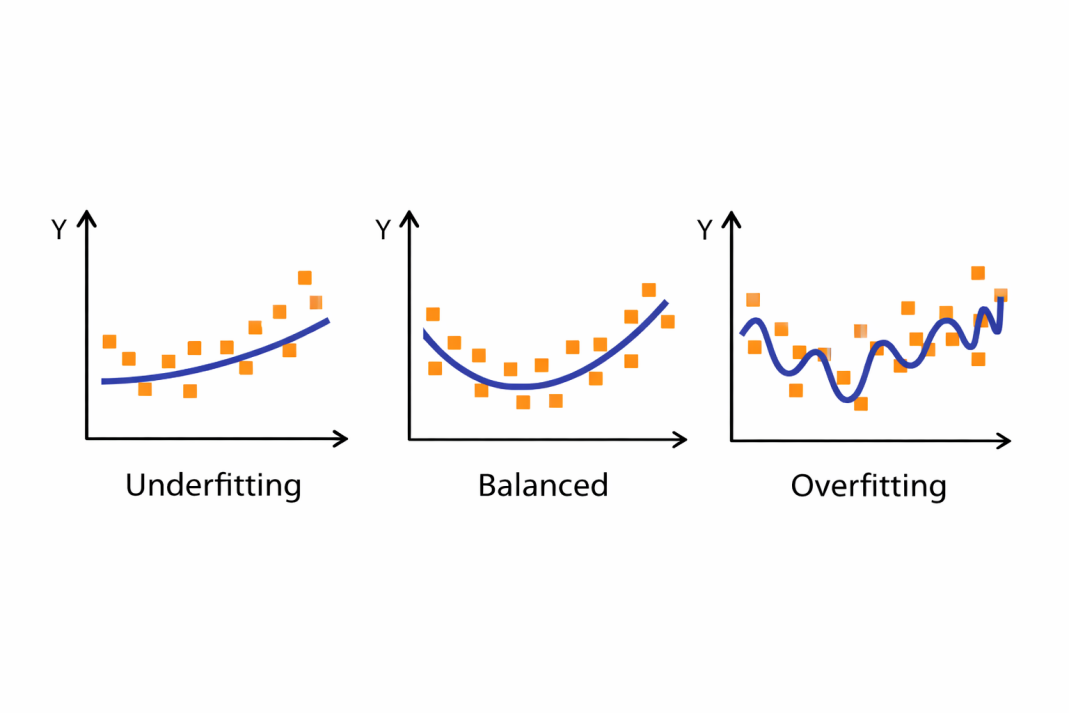
Principal Component Analysis (PCA) (6)

- Since eigenvalue equals the variance of the principal component, we can decide how many components we would use from the data
  - You can see that at which point adding the dimensions do not help explaining the variance in the data.



Regularization: Motivation (1)

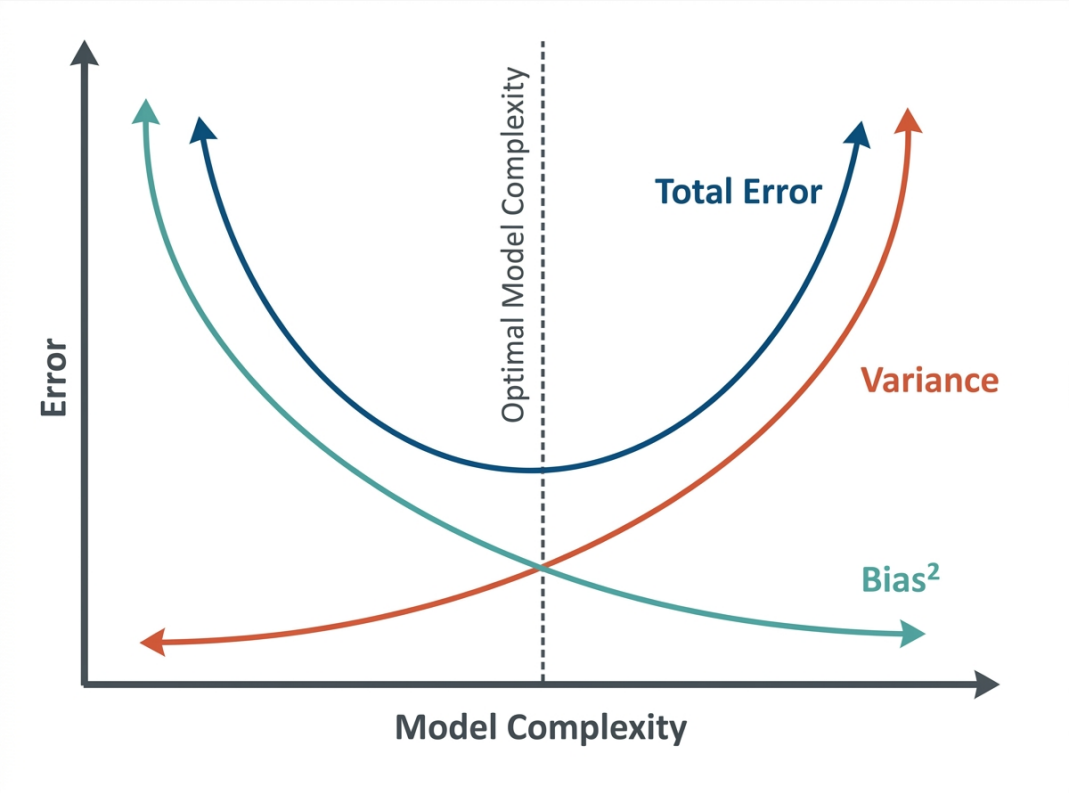
- We want to make the function flexible enough to minimize the bias
- However, we do not want the overfitting



- As figure illustrates, overfitting is caused by high variance
  - There is a trade-off: underfitting happens because of restrictive model, whereas overfitting happens because of complex model

Regularization: Motivation (2)

- Last time, we learn that to minimize the out-of-sample prediction error, we want to minimize both bias and variance
  - Only minimizing bias might not be optimal for variane



- Regularization introduces bias by limiting the model complexities, but shrinks variance

Ridge regression (1)

- Think about constraining the model complexity for linear model by adding constraints on coefficients

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} (Y_i - \alpha - \boldsymbol{X}_i^\top \beta)^2 \quad \text{s.t.} \quad ||\beta||_2^2 \leq t$$

  - Note that $||\beta||_2^2 = \sum_{j=1}^{p} \beta_j^2$

- Using Lagrangian multiplier, the dual problem is written as

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} (Y_i - \alpha - \boldsymbol{X}_i^\top \beta)^2 + \lambda(||\beta||_2^2 - t)$$

  with $\lambda \geq 0$

  - For a fixed λ, the term $-\lambda t$ is constant, so it is equal to minimizing[1]

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} (Y_i - \alpha - \boldsymbol{X}_i^\top \beta)^2 + \lambda||\beta||_2^2$$

[1] Notice that we do not optimize the Lagrangian multiplier λ. So, the equivalence here is that there exists a correspondence between λ and t and this is one-to-one, but the mapping is specific to each data set.

Ridge regression (2)

- Now, let $\tilde{Y}_i = Y_i - \bar{Y}$ and $\tilde{\boldsymbol{X}}_i = \boldsymbol{X}_i - \bar{\boldsymbol{X}}$, and the model becomes

$$\tilde{Y}_i = \tilde{\boldsymbol{X}}_i^\top \beta + \epsilon_i$$

  - By standardization, we do not have an intercept
  - Then, optimization problem is

$$\hat{\beta} = \arg\min_{\beta} ||\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\beta||_2^2 + \lambda||\beta||_2^2$$

- By taking the derivative, we get

$$\begin{aligned}
&\frac{\partial}{\partial \beta}\left(||\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\beta||_2^2 + \lambda||\beta||_2^2\right) \\
&= -2\tilde{\boldsymbol{X}}^\top(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\beta) + 2\lambda\beta = 0
\end{aligned}$$

- This gives us the closed-form solution

$$\hat{\beta} = (\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + \lambda \boldsymbol{I})^{-1} \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{Y}}$$

Ridge regression (3)

- When $\lambda > 0$, $\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + \lambda \boldsymbol{I}$ is always invertible

- Recall that matrix is invertible if it is positive definite
  - Now, consider the quadratic form

$$\boldsymbol{v}^\top (\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + \lambda \boldsymbol{I}) \boldsymbol{v} = ||\tilde{\boldsymbol{X}}^\top \boldsymbol{v}||_2^2 + \lambda ||\boldsymbol{v}||_2^2$$

  - Now, even if $||\tilde{\boldsymbol{X}}^\top \boldsymbol{v}||_2^2 = 0$, $||\boldsymbol{v}||_2^2 > 0$ as $\boldsymbol{v} \neq \boldsymbol{0}$
  - So, when $\lambda > 0$, $\boldsymbol{v}^\top (\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + \lambda \boldsymbol{I}) \boldsymbol{v} > 0$, meaning that $\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + \lambda \boldsymbol{I}$ is always invertible

- This is different from OLS
  - In OLS, $\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}$ can be not invertible when regressors are rank-deficient (or known as perfect collinearity)
  - This is likely when you have so many regressors
  - Ridge regression solves this problem

Ridge regression (4)

- However, it is important to recognize that the coefficient of ridge regression is biased!

- To see why,

$$\begin{aligned}\mathbb{E}[\hat{\beta} \mid \boldsymbol{X}] &= (\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + \lambda \boldsymbol{I})^{-1} \tilde{\boldsymbol{X}}^\top \mathbb{E}\,\tilde{\boldsymbol{Y}} \mid \boldsymbol{X}] \\ &= (\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + \lambda \boldsymbol{I})^{-1} \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} \beta\end{aligned}$$

  which is different from β unless $\lambda = 0$

- This makes sense because regularization makes the variance smaller but introduces the bias (called regularization bias)
  - If you need to use models with regularization for statistical inference, you need to make your estimation robust to regularization bias
  - This is exactly what double machine learning is doing.

Ridge regression (5)

- Recall that SVD gives you $\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$
  - Here, $\boldsymbol{U} \in \mathbb{R}^{n \times r}$, $\boldsymbol{V} \in \mathbb{R}^{p \times r}$m and $\boldsymbol{D} = \mathrm{diag}(d_1, \cdots, d_r)$ where $r = \mathrm{rank}(\tilde{\boldsymbol{X}})$.

- Plugging in the closed-form solution for ridge, we get

$$\begin{aligned}\hat{\beta} &= (\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + \lambda \boldsymbol{I})^{-1} \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{Y}} \\ &= (\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top + \lambda \boldsymbol{I})^{-1} \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top \tilde{\boldsymbol{Y}} \\ &= \boldsymbol{V}(\boldsymbol{D}^2 + \lambda \boldsymbol{I})\boldsymbol{V}^\top \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top \tilde{\boldsymbol{Y}}\end{aligned}$$

  where the last equality is because $\boldsymbol{V}$ is orthonormal.

- Thus, let $\boldsymbol{z} = \boldsymbol{U}^\top \tilde{\boldsymbol{Y}}$. Then,

$$\hat{\beta} = \sum_{j=1}^{r} \frac{d_j}{d_j^2 + r} z_j \boldsymbol{v}_j$$

  - Notice that this formulation does not require you to calculate the inverse (i.e., computationally efficient)
  - Similar logic to QR decomposition to OLS

Added: Bias-Variance Tradeoff of Ridge

- Recall that we derived

$$\hat{\beta} = (\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + \lambda \boldsymbol{I})^{-1} \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{Y}}$$

  - For the sake of simplicity, homoskedasticity $\mathbb{V}[\epsilon \mid \boldsymbol{X}] = \sigma^2 \boldsymbol{I}_n$

- Then, the variance is written as

$$\begin{aligned}
\mathbb{V}(\hat{\beta} \mid \tilde{\boldsymbol{X}}) &= \mathbb{V}((\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + \lambda \boldsymbol{I})^{-1} \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{Y}} \mid \tilde{\boldsymbol{X}}) \\
&= \mathbb{V}((\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + \lambda \boldsymbol{I})^{-1} \tilde{\boldsymbol{X}}^\top [\tilde{\boldsymbol{X}}\beta + \boldsymbol{\epsilon}] \mid \tilde{\boldsymbol{X}}) \\
&= \mathbb{V}((\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + \lambda \boldsymbol{I})^{-1} \tilde{\boldsymbol{X}}^\top \boldsymbol{\epsilon} \mid \tilde{\boldsymbol{X}}) \\
&= (\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + \lambda \boldsymbol{I})^{-1} \tilde{\boldsymbol{X}}^\top \mathbb{V}(\boldsymbol{\epsilon} \mid \tilde{\boldsymbol{X}}) \tilde{\boldsymbol{X}} (\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + \lambda \boldsymbol{I})^{-\top} \\
&= \sigma^2 (\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + \lambda \boldsymbol{I})^{-1} \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} (\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + \lambda \boldsymbol{I})^{-\top}
\end{aligned}$$

- Notice that as λ increases, the variance decreases
  - This is the trade-off between bias and variance in ridge regression

Lasso regression (1)

- Instead of using L2 norm $||\boldsymbol{\beta}||_2^2$ for penalty, consider

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \arg\min_{\alpha,\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - \alpha - \boldsymbol{X}_i^\top \boldsymbol{\beta})^2 + \lambda ||\boldsymbol{\beta}||_1$$

  then $||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p} |\beta_j|$.

- As with the previous case, the formulation above is dual problem of

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \arg\min_{\alpha,\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - \alpha - \boldsymbol{X}_i^\top \boldsymbol{\beta})^2 \quad \text{s.t.} \quad ||\boldsymbol{\beta}||_1 \leq t$$

Lasso regression (2)

- Lasso regression makes the coefficient tend to be 0, whereas Ridge regression makes the coefficient smaller
  - Intuition can be seen from this figure (try to understand!)